Synthetic Health Data: The Good, the Bad, and the Ugly

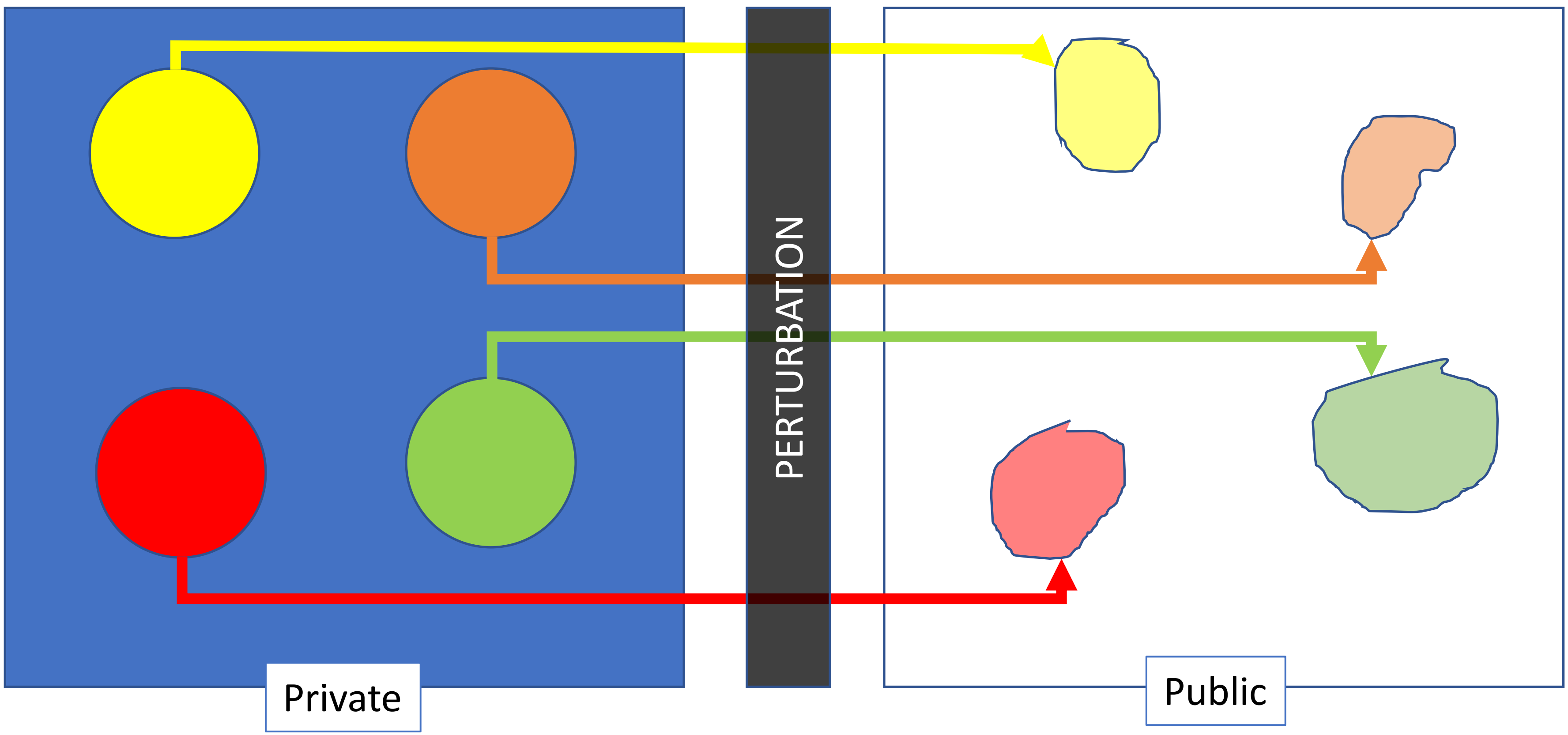
Bradley Malin, Ph.D.

Accenture Professor of Biomedical Informatics, Biostatistics, & Computer Science
Co-Director, Center for Health Data Science
Co-Director, Center for Genetic Privacy & Identity in Community Settings
Co-Director, Big Biomedical Data Science Training Program

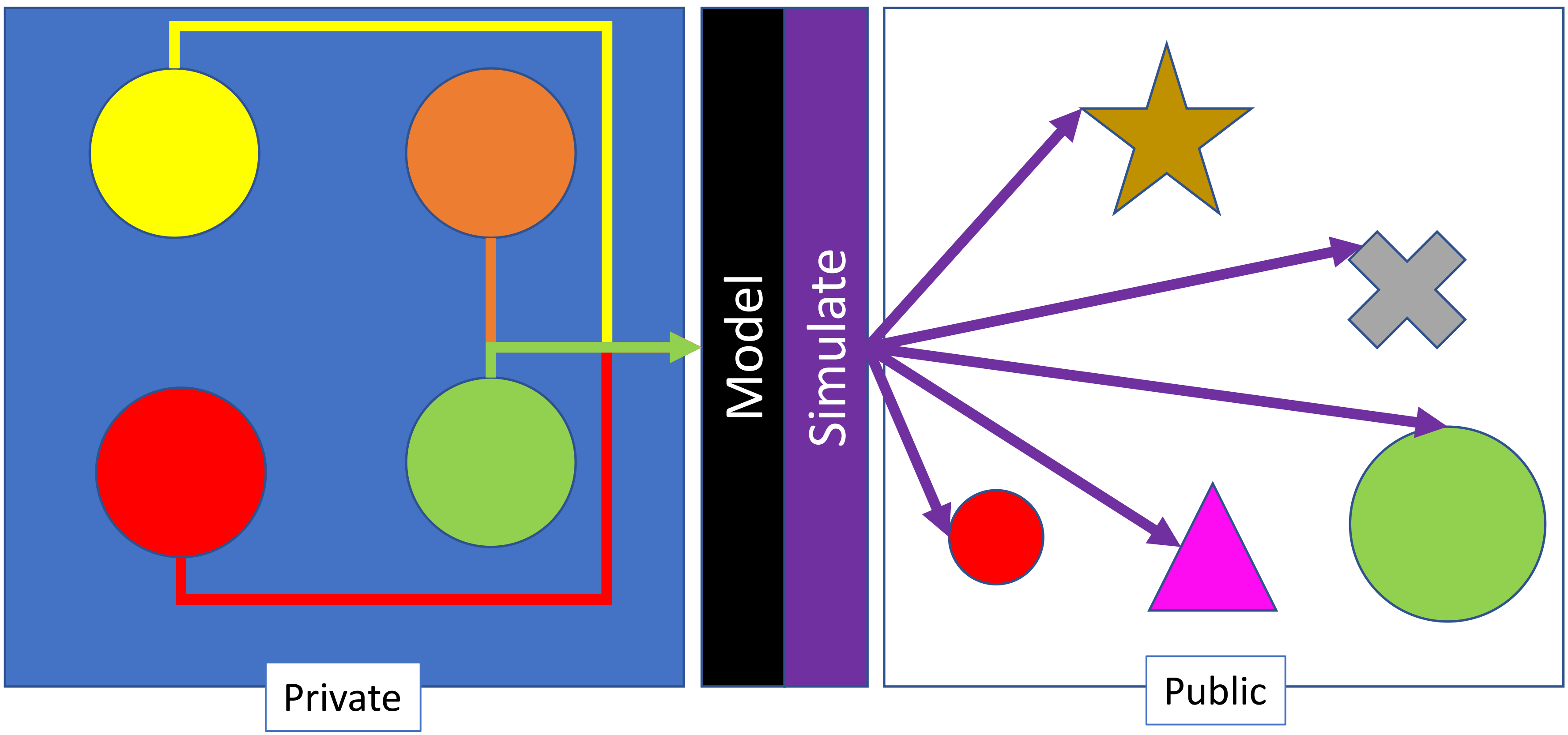
VANDERBILT UNIVERSITY
MEDICAL CENTER

November 10, 2021

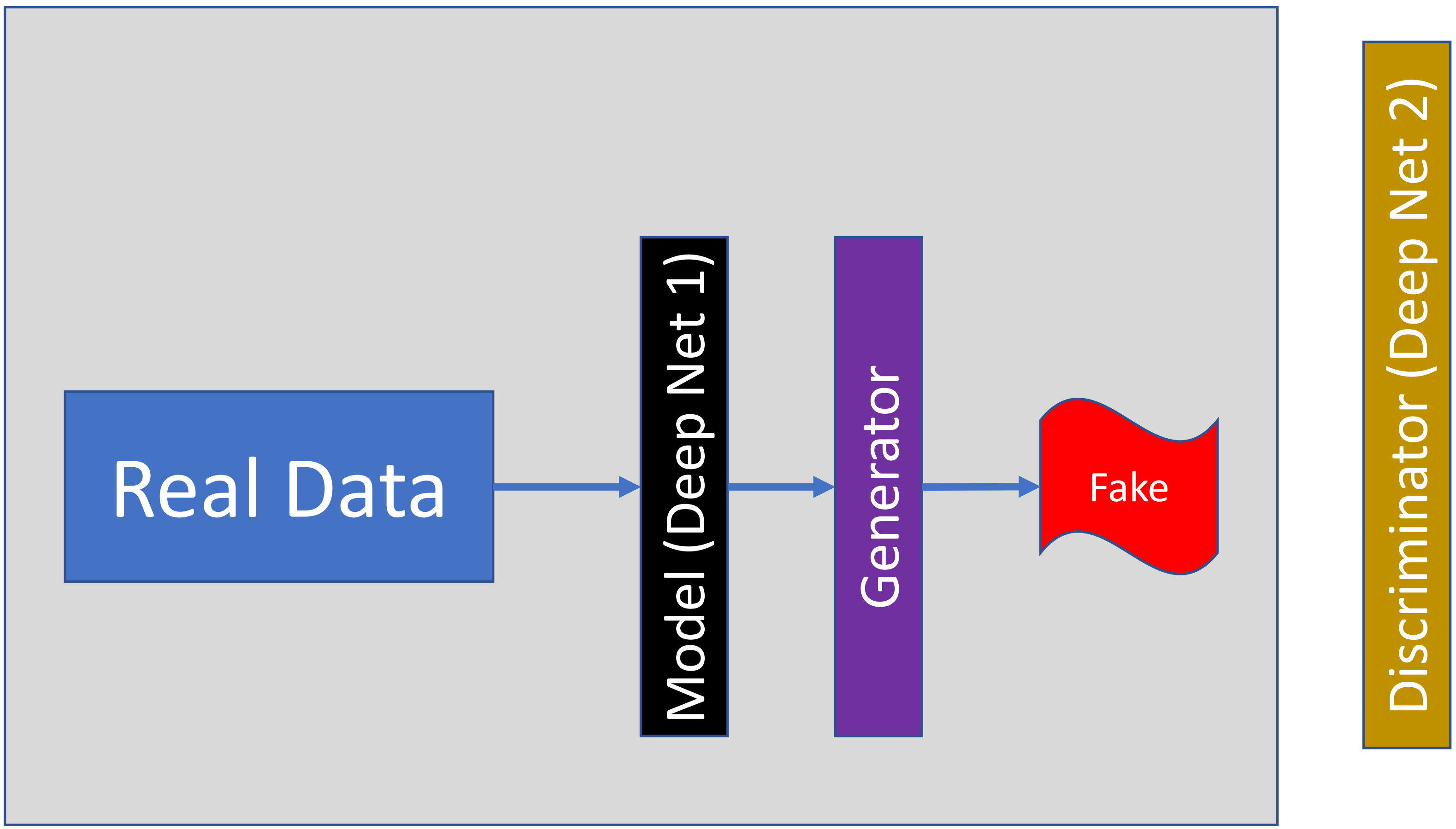
Ways to Generate Synthetic Data: Perturbation



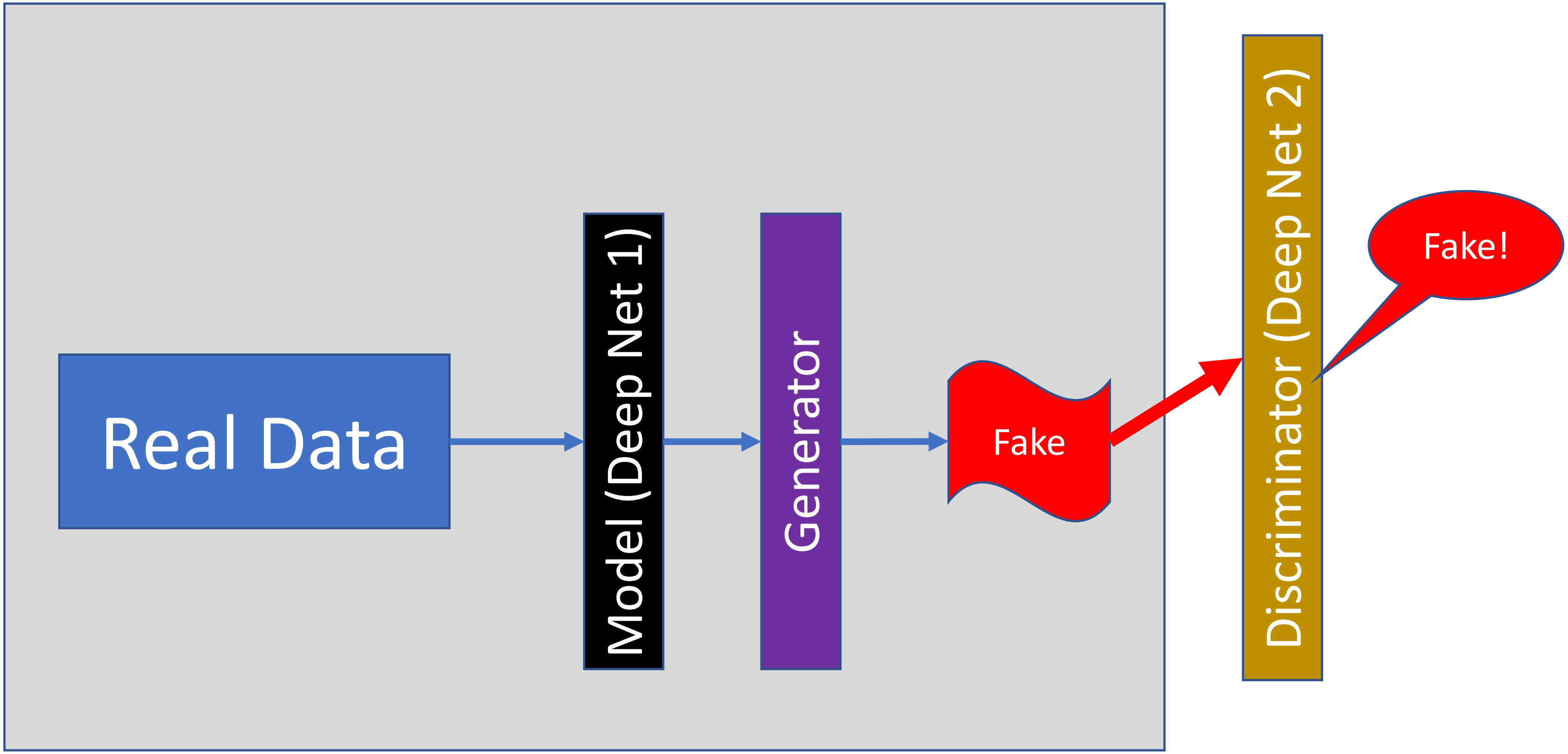
Ways to Generate Synthetic Data: Simulation



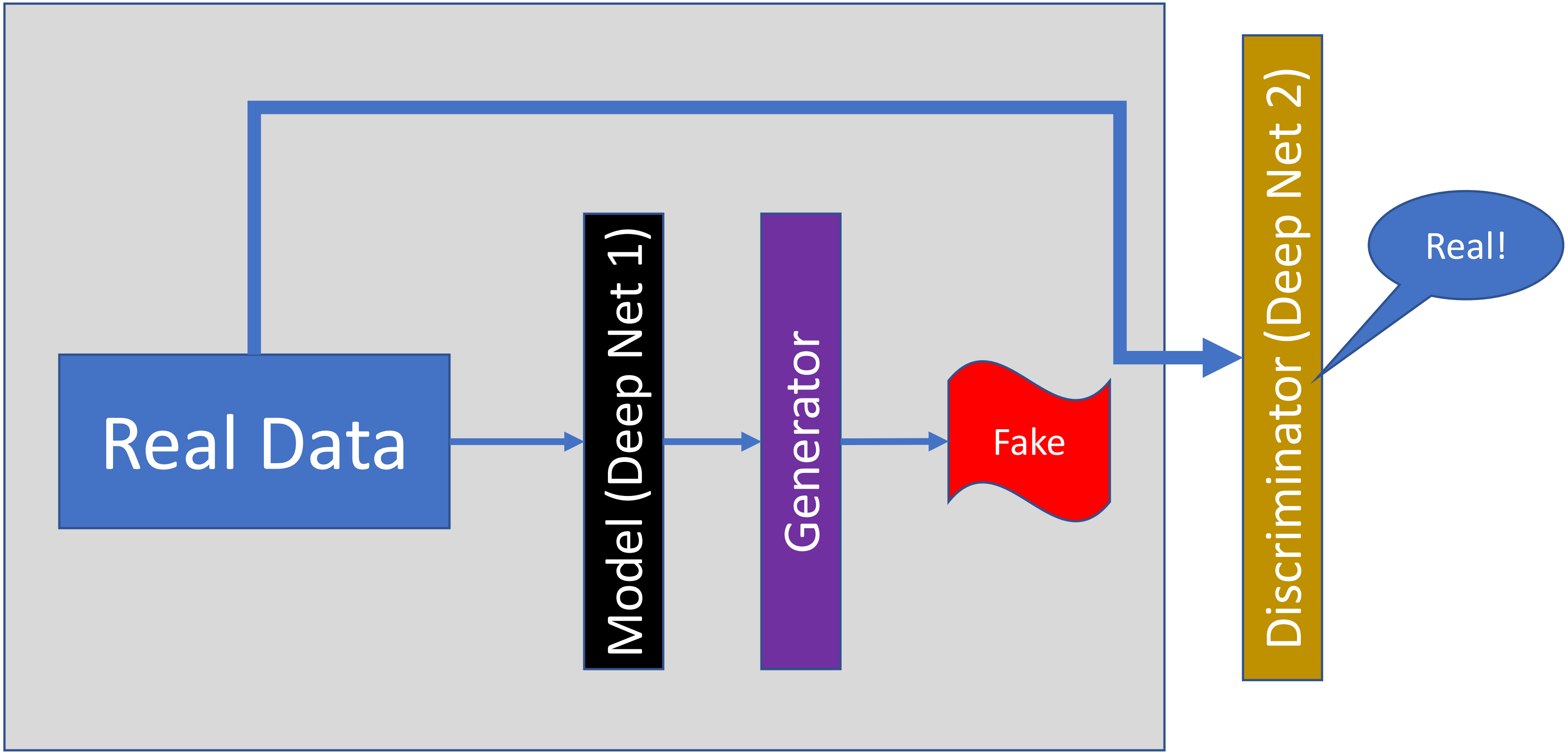
Generative Adversarial Networks: GANs



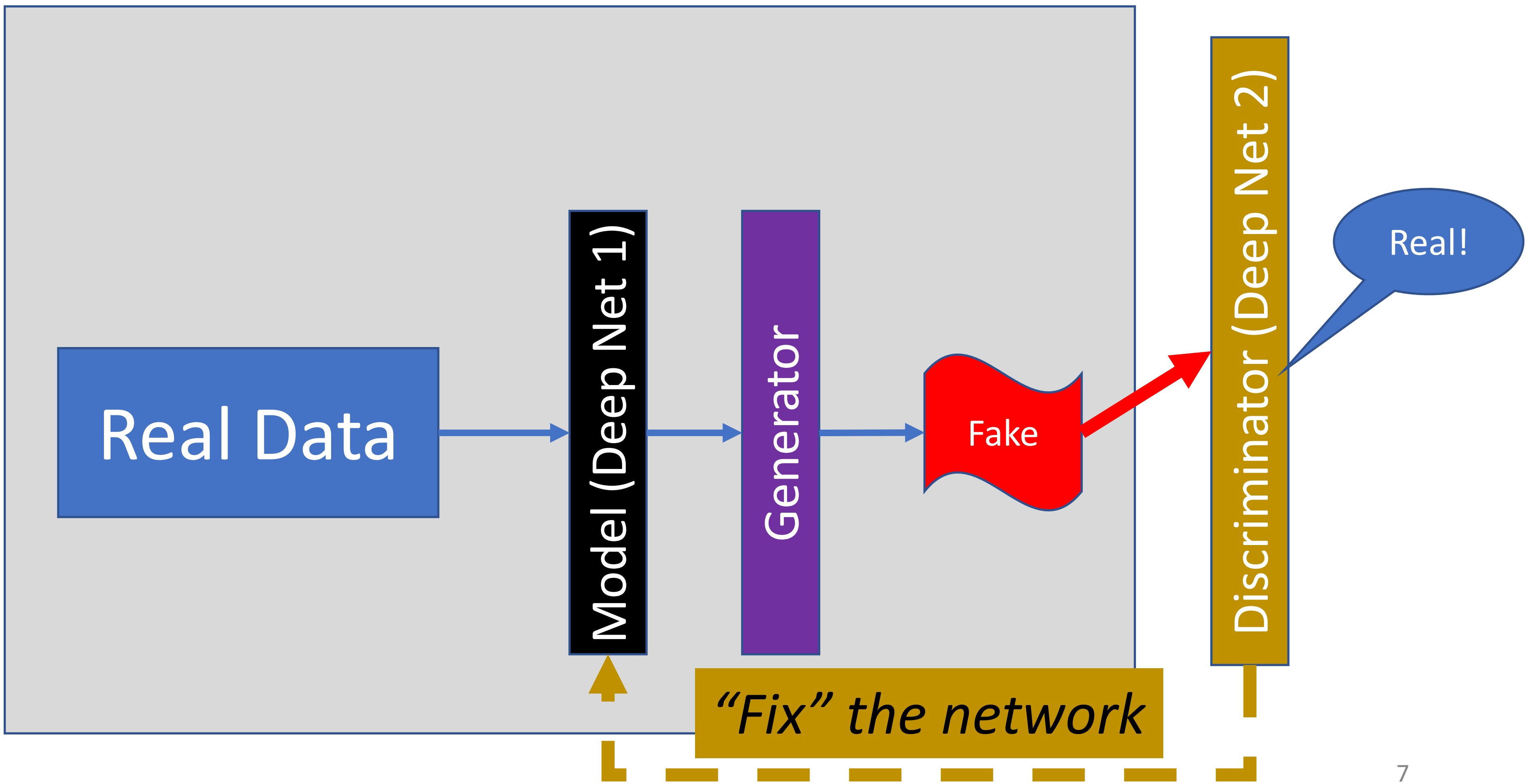
Generative Adversarial Networks: GANs



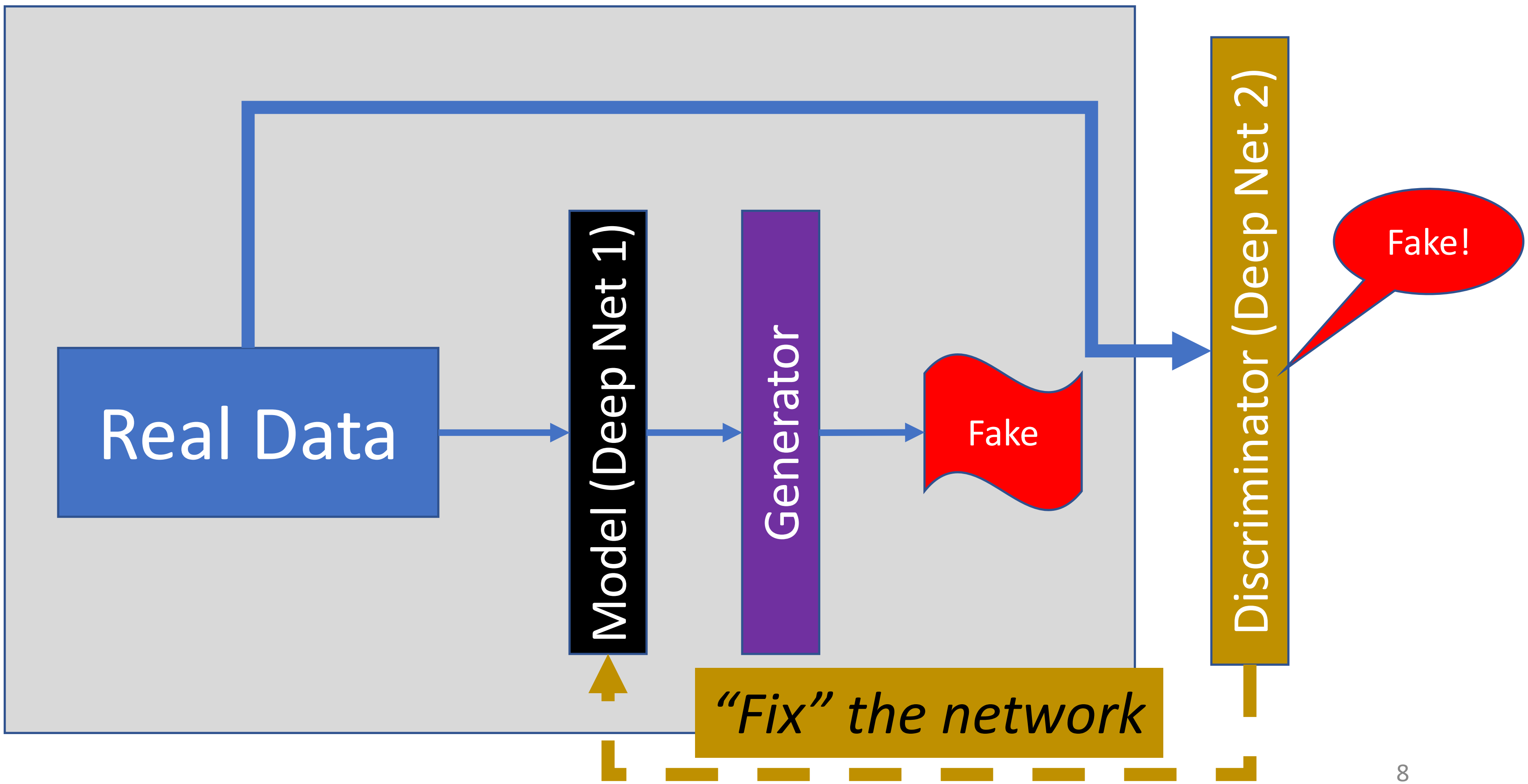
Generative Adversarial Networks: GANs



Playing the GAN Game



Playing the GAN Game



This is Not a New Principle

Ian Goodfellow
@goodfellow_ian

4.5 years of GAN progress on face generation.
arxiv.org/abs/1406.2661 arxiv.org/abs/1511.06434
arxiv.org/abs/1606.07536 arxiv.org/abs/1710.10196
arxiv.org/abs/1812.04948

Journal of Official Statistics, Vol. 18, No. 4, 2002, pp. 531–543

Satisfying Disclosure Restrictions With Synthetic Data Sets

Jerome P. Reiter[1]

To avoid disclosures, Rubin proposed
so that (i) no unit in the released data
and (ii) statistical procedures that are
In this article, I show through simu
from synthetic data in a variety of se
proportional to size sampling, two-s
provide guidance on specifying the n
the benefit of including design variab

Key words: Confidentiality; disclosur

JPrivacy Confidentiality

Current Archives Announcements TPDP workshop Submissions

Home / Archives / Vol. 1 No. 1 (2009): Inaugural Issue / **Articles**

Estimating Risks of Identification Disclosure in Partially Synthetic Data

PDF

Published: Apr 1, 2009

DOI: https://doi.org/10.29012/jpc.v1i1.567

Keywords:
Confidentiality Public use data

Jerome P. Reiter
Department of Statistical Science, Duke University, Durham, NC
https://orcid.org/0000-0002-8374-3832

Robin Mitra
University of Southampton, Southampt
https://orcid.org/0000-0001-9584

Abstract

To limit disclosures, statistical agencies

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 583–590

Inferentially Valid, Partially Synthetic Data: Generating from Posterior Predictive Distributions not Necessary

Jerome P. Reiter[1] *and Satkartar K. Kinney*[2]

To avoid disclosures in public use microdata, one approach is to release partially synthetic data sets. These comprise the units originally surveyed with some collected values, for example sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. In practice, partially synthetic data typically are generated from Bayesian posterior predictive distributions; that is, one draws repeated values of parameters in the synthesis models before generating data from them. We show, however, that inferentially valid, partially synthetic data can be generated by fixing the parameters of the synthesis models at their modes. We do so with both a theoretical example and illustrative simulation studies. We also discuss implications of these results for agencies generating synthetic data.

Key words: Confidentiality; disclosure; imputation; microdata; privacy; survey.

This is Not a New Principle
(Choi MLHC 2017)

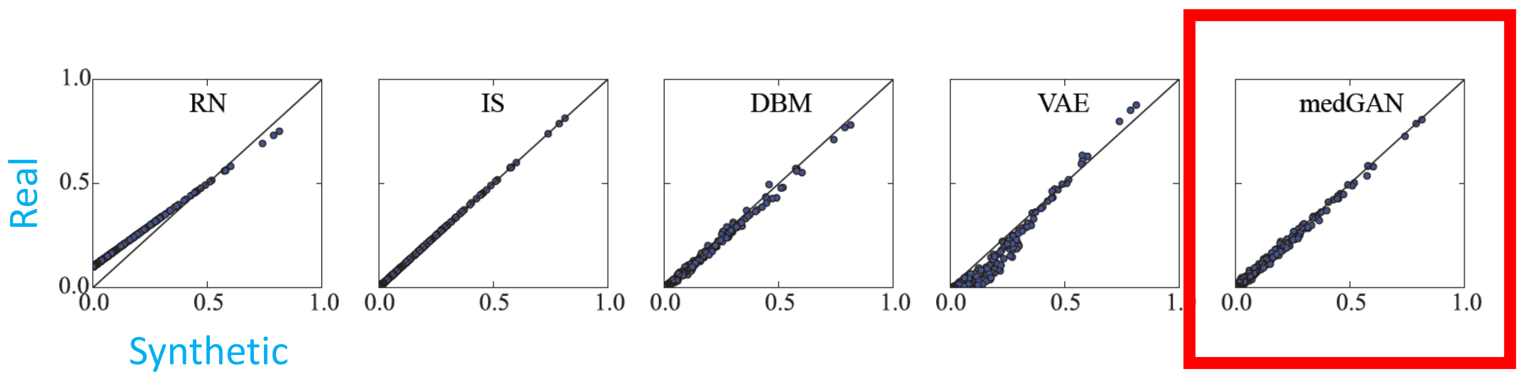
Proceedings of Machine Learning for Healthcare 2017 JMLR W&C Track Volume 68

Generating Multi-label Discrete Patient Records using Generative Adversarial Networks

Edward Choi[1] MP2893@GATECH.EDU
Siddharth Biswal[1] SBISWAL7@GATECH.EDU
Bradley Malin[2] BRADLEY.MALIN@VANDERBILT.EDU
Jon Duke[1] JON.DUKE@GATECH.EDU
Walter F. Stewart[3] STEWARWF@SUTTERHEALTH.ORG
Jimeng Sun[1] JSUN@CC.GATECH.EDU

[1]Georgia Institute of Technology [2] Vanderbilt University [3] Sutter Health

- Sutter Health & MIMIC
- Demographics, Diagnoses, Procedures, & Meds
- Prediction of presence / absence clinical concept



Limitations

- Autoencoder induced noise and hurt learning

- Evaluation measures based on superficial aspects of data gave false impression of merits of simulation

- Focus on all EHR data led to overrepresentation of common associations

Evolution

- Better training (Wasserstein distance) and evaluation methods (latent dimensions) (Zhang JAMIA 2020)

- Enabling constraints (e.g., preventing women from having prostate cancer) (Yan AMIA 2020)

- Move from static to longitudinal data: think LSTMs + GANs (Zhang JAMIA 2021)

Zhang, Yan, Mesa, Sun, & Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. JAMIA. 2020; 27: 99-108.
Yan, Zhang, Nyemba, & Malin. Generating electronic health records with multiple data types and constraints. Proc AMIA Symp. 2020: 1335-1344.
Zhang, Yan, Lasko, Sun, & Malin. SynTEG: A framework for temporal structured electronic health data simulation. JAMIA. 2021; 28: 596-604.

Building a Synthetic Resource

Data selection

235,000 participants w/
demographics, physical measures, survey responses

Model training

Train multiple **GAN** models
to determine the best converge point and control repetition

Based on **specific requirements**
(e.g., volume, constraints)

Data generation

Reinsert important outliers that are discarded by
the data generation model, e.g. Wolfram syndrome

Post-hoc curation

System/software Development

- Develop data analytic tools
- Test important system features
- Complete quality control and assurance tasks

All of Us RESEARCH PROGRAM

Case Study for Demos & Tutorial

> 30 researcher outreach and training events

> 2000 users

Researcher Workbench launched

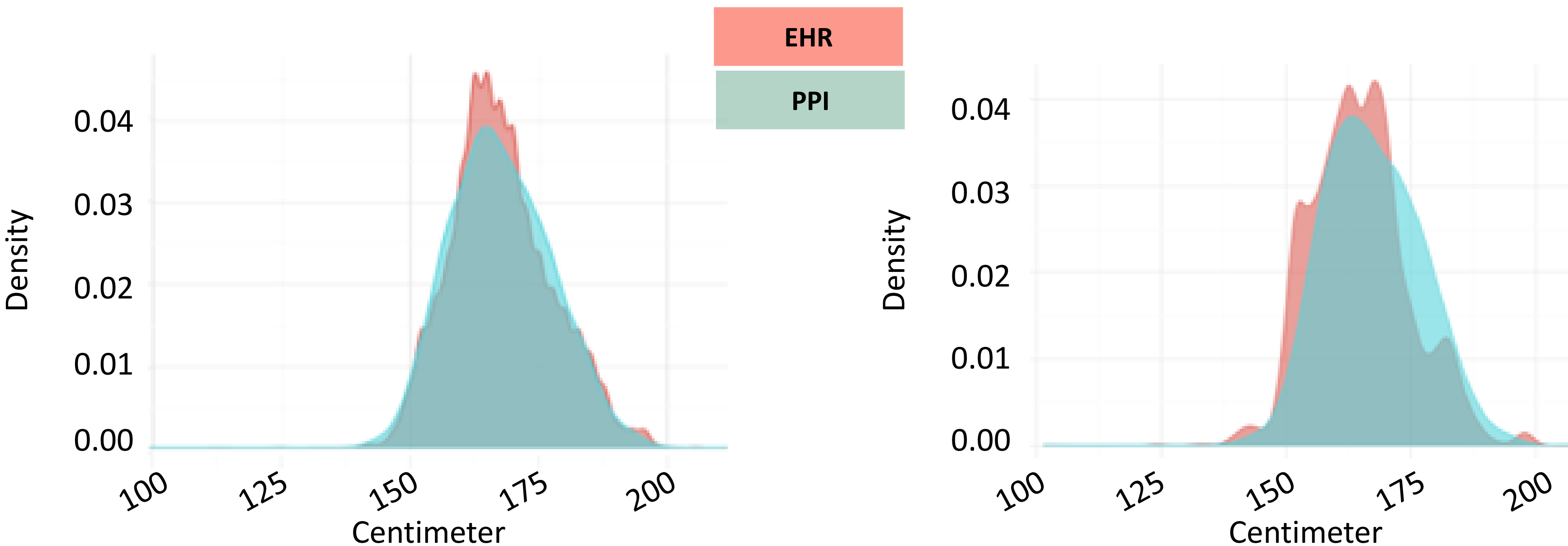
1 year after Launch

May 2020

May 2021

All of Us RESEARCH PROGRAM

Real *vs* Synthetic in the Same Tutorial



Using ***real*** data in RW

Using ***synthetic*** data in mirror RW

All of Us RESEARCH PROGRAM

Is Synthetic Data "De-identified"?

<u>According to HIPAA (Privacy Rule)</u>:

"information that does not identify an individual and … no reasonable basis … information can be used to identify an individual"

Safe Harbor

- Removal of 18 types of identifiers
- No actual knowledge residual information can identify individual

Expert Determination

- Apply statistical or scientific principles
- Very small risk that anticipated recipient could identify individual

What Could Go Wrong?

ARTIFICIAL INTELLIGENCE

AI fake-face generators can be rewound to reveal the real faces they trained on

Researchers are calling into doubt the popular idea that deep-learning models are "black boxes" that reveal nothing about what goes on inside

By Will Douglas Heaven

October 12, 2021

https://arxiv.org/abs/2107.06304

Deep Neural Networks are Surprisingly Reversible: A Baseline for Zero-Shot Inversion

Xin Dong[1,2], Hongxu Yin[1], Jose M. Alvarez[1], Jan Kautz[1], and Pavlo Molchanov[1]
[1]NVIDIA, [2]Harvard University
xindong@g.harvard.edu, {dannyy, josea, pmolchanov, jkautz}@nvidia.com

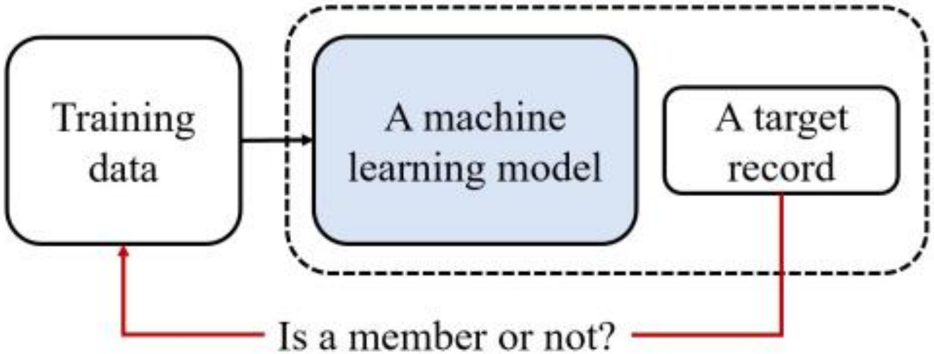
FAKE

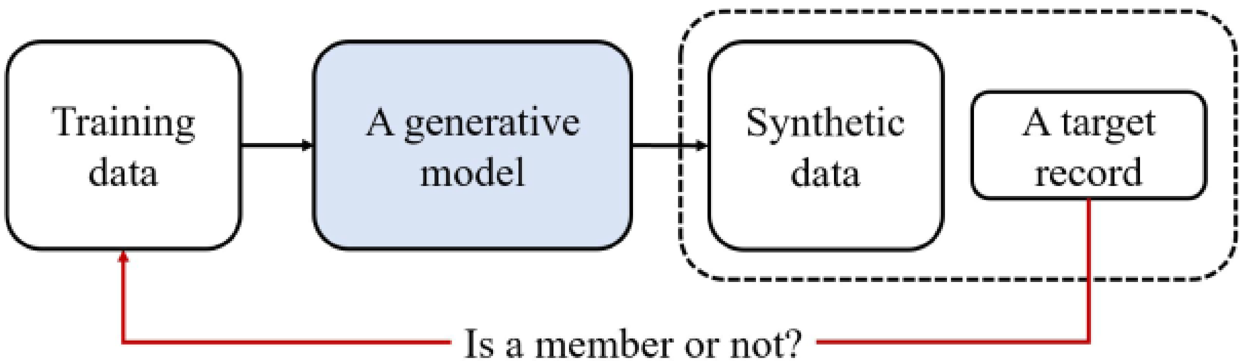
Real

A Bunch of Things

- Mimic
  - Insufficient training data can lead to “mimicking” of original records
- Membership Inference
  - User can test if features of someone they know appear to be in the training data
  - Requires knowing the features in question
- Attribute Inference
  - User can predict features (they don’t know) about someone based on features they do know
- Combining Membership and Attribute is where disclosure occurs

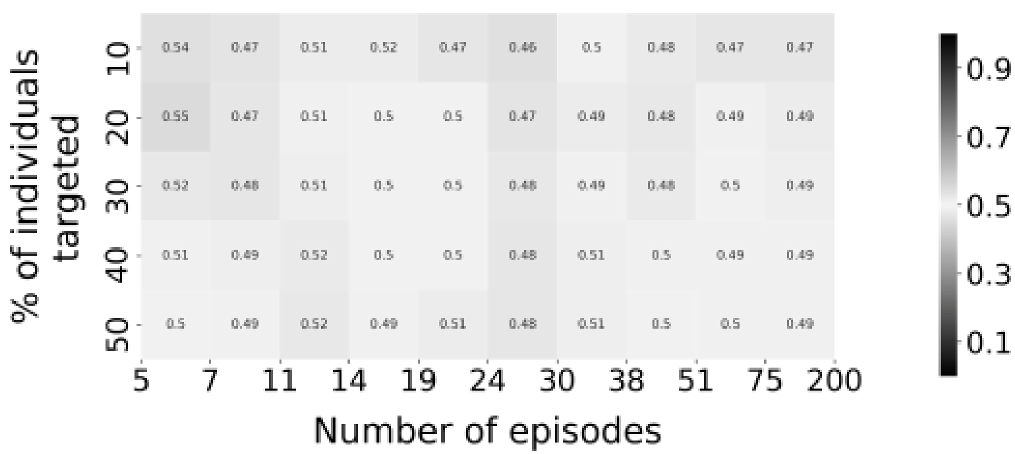
Membership Intrusion



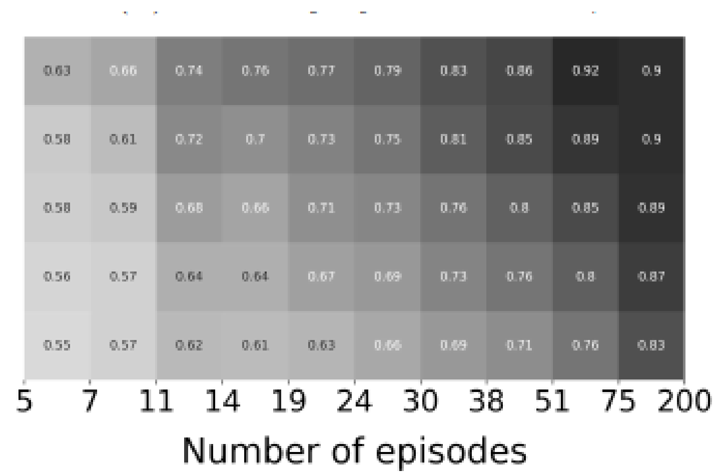


An Attack on VUMC Data

- 45,000 patients, diagnosis and procedure codes
- Up to 200 visits
- Adversary has 10% “prior” knowledge



Fully Synthetic



Partially Synthetic

Context Matters ALOT

- Must define the expected capabilities of the recipients of the data
- Privacy assessments should consider the data, as well as how the data was created
- Must consider the recipient’s tolerance for errors
- Most consider society’s tolerance for intrusion (and claimed intrusion)

Questions?

b.malin@vanderbilt.edu

Center for Genetic Privacy & Identity in Community Settings
https://www.vumc.org/getprecise

Health Data Science Center
https://www.vumc.org/heads